Analysis of cross-cultural comparability of PISA 2009 scores

Kankaras, M.; Moors, G.B.D.

Published in:
Journal of Cross-Cultural Psychology

DOI:
10.1177/0022022113511297

Publication date:
2014

Document Version
Publisher's PDF, also known as Version of record

Analysis of Cross-Cultural Comparability of PISA 2009 Scores

Milos Kankaras and Guy Moors

Journal of Cross-Cultural Psychology 2014 45: 381 originally published online 15 November 2013
DOI: 10.1177/0022022113511297

Journal of Cross-Cultural Psychology
2014, Vol. 45(3) 381–399

DOI: 10.1177/0022022113511297
jccp.sagepub.com
SAGE

Analysis of Cross-Cultural Comparability of PISA 2009 Scores

Miloš Kankaraš[1] and Guy Moors[1]

Abstract

The Program for International Student Assessment (PISA) is a large-scale cross-national study that measures academic competencies of 15-year-old students in mathematics, reading, and science from more than 50 countries/economies around the world. PISA results are usually aggregated and presented in so-called "league tables," in which countries are compared and ranked in each of the three scales. However, to compare results obtained from different groups/countries, one must first be sure that the tests measure the same competencies in all cultures. In this paper, this is tested by examining the level of measurement equivalence in the 2009 PISA data set using an item response theory approach (IRT) and analyzing differential item functioning (DIF). Measurement in-equivalence was found in the form of uniform DIF. In-equivalence occurred in a majority of test questions in all three scales researched and is, on average, of moderate size. It varies considerably both across items and across countries. When this uniform DIF is accounted for in the in-equivalent model, the resulting country scores change considerably in the cases of the "Mathematics," "Science," and especially, "Reading" scale. These changes tend to occur simultaneously and in the same direction in groups of regional countries. The most affected seems to be Southeast Asian countries/territories whose scores, although among the highest in the initial, homogeneous model, additionally increase when accounting for in-equivalence in the scales.

Keywords

measurement equivalence, PISA, differential item functioning, cross-cultural research, educational measurement

Introduction

A large-scale comparative study "The Programme for International Student Assessment" (PISA) measures competencies of 15-year-old students in mathematics, reading, and science in more than 50 countries around the world. The results of these tests are transferred into average country scores and presented in a number of "league tables," in which countries are ranked in each of the three competencies. A new set of results from the last round of PISA testing conducted in 2009 are presented (OECD, 2010). Regarded to be one of the best comparative studies up to date,

[1]Tilburg University, Netherlands

Corresponding Author:
Miloš Kankaraš, Department of Methodology and Statistics, Tilburg University, Room P 1.109, P.O. Box 90153, 5000 LE Tilburg, Netherlands.
Email: m.kankaras@tilburguniversity.edu

during the last 10 years, PISA has had enormous impact on scholarly community, on mass-media and public, and on national and international policy makers.

Although the results have been widely discussed and controversy erupted about possible causes of differences in country scores, the underlying assumption that tests are comparable across all countries has rarely been questioned. However, because of the cultural and linguistic diversity in participating countries, PISA tests may not be equivalent in all countries, which would make comparisons unfair. For example, instead of measuring only academic competencies, test results may be influenced by translation, different familiarity of testing format, cultural bias in content of questions, and so on. If this is the case, comparison of country scores is questionable at best. Therefore, before comparing results across countries, it is crucial to verify measurement equivalence, that is "whether or not, under different conditions of observing and studying phenomena, measurement operations yield measures of the same attribute" (Horn & McArdle, 1992, p. 177).

The aim of this work is to analyze measurement equivalence of PISA 2009 scales. First, we examine whether and to what extent measurement in-equivalence is present in the PISA 2009 scales. Second, in case of in-equivalence, we want to determine the consequences it has on the country scores and consecutive country rankings. For this purpose, we use the item response theory (IRT) approach in analyzing measurement equivalence, namely, the analysis of differential item functioning (DIF).

The remainder of this article is organized as follows. First, we elaborate on the issue of measurement equivalence and its relevance in relation to the PISA surveys. After introducing the data and methodological approach, results are presented and discussed. We will show that at least part of the country rankings of PISA scores are biased by lack of recognition of the issue of measurement in-equivalence.

Comparing PISA Country Ratings: A Fundamental Assumption

Launched in 1997, PISA is conducted in 3-yearly cycles and examines reading literacy, mathematical, and scientific competency (and, in 2003, problem solving) of national samples of 15-year-old students to "allow national policymakers to compare the performance of their education systems with those of other countries" (OECD, 1999, p. 7; www.pisa.oecd.org). As a consequence of this intention, most of official reports are communicated in the form of "league tables" that present country rankings in each of the three subject areas (as well as in subscales of the subject that was in focus in a particular cycle). In PISA 2009, countries were ranked in the three "overall" scales and in five reading subscales with Shanghai-China finishing first in all but one reading subscale, and Korea, Finland, and Hong Kong following closely. These rankings initiate substantial media and public attention and a "flood" of discussions, and have major impact not only on policy-making, but also on the public consciousness (Fertig, 2004; Mortimore, 2009; Pongratz, 2006). But the question is whether these "league tables" are valid representatives of country scores in measured competencies?

It is generally accepted that tests used in a cross-cultural study should be culturally fair, which means that two students with the same level of ability will get the same scores irrespective of the country they are coming from (Hui & Triandis, 1985; Poortinga, 1989; see Hambleton, Merenda, & Spielberger, 2005—Standards for Educational and Psychological Testing). However, there are a number of potential factors that might hamper the comparability in results. One such issue that emerges is that part of the equivalence of a test instrument is almost always lost when translating it into other languages because each translated word or sentence necessary alters the meaning of the original formulation to a certain degree (Bonnet, 2002; Grisay & Monseur, 2007; Harkness, Pennell, & Schoua-Glusberg, 2004; Sjøberg, 2007). But even if a perfect translation were possible, in addition to language, there are several other factors that should be taken into consideration

if scores on tests that have been adapted for use in multiple languages and cultures are to be meaningfully interpreted. These include item content and format familiarity, curricula, examinee motivation, economic status, standard of living, cultural values, test administration conditions, test anxiety, response biases, test speediness, and so on (Hambleton & Kanjee, 1993; Hui & Triandis, 1989; van de Vijver & Poortinga, 1991). Whenever any of these factors affects test scores differently across cultural groups, comparability of country scores will be hindered because they will be measuring different phenomena (Hui & Triandis, 1985; van de Vijver, 1998). Hence, to make valid comparisons of PISA country scores, it is necessary to establish measurement equivalence ("measurement invariance") or avoid DIF within the context of IRT (Hambleton et al., 2005; Kankaraš & Moors, 2010; Mellenbergh, 1989).

Although it is claimed that PISA provides "strong, cross-culturally valid measures of competencies" (OECD, 2001, p. 27), in the initial years of PISA project, this cross-cultural validity of PISA scales has not been thoroughly investigated (Allerup, 2007; Bechger, 2006; Bonnet, 2002; Goldstein, 2008). This comparability is further questioned by studies that have found in-equivalence in (parts of) PISA data sets. Goldstein (2008) established evidence for differential loadings when comparing scores of France and England. Yildirim (2006) showed that factor structures of PISA results were not equivalent across Turkish and American samples. Allerup (2007) demonstrated that the assumption of temporal equivalence between individual cycles from 2003 and 2006 waves is not valid for the Danish sample. Grisay and Monseur (2007) found that proportion of in-equivalent test items in PISA 2000 reading scale is consistently around 25% to 30%.

These studies clearly indicate that the issue of measurement equivalence may be problematic in the PISA study. Given that the PISA project is regarded as a highly valuable comparative study, testing more than a million of students in more than 50 countries of the world, it comes as a surprise that relatively little systematic inquiry into cross-cultural equivalence in measurement has been conducted. Taking into account the impact of the PISA project, there is urgent need for this. With new results from the 2009 wave recently released (December 2010), a new round of country comparisons has began again and the basic, fundamental question of the validity of these comparisons needs to be tackled. This is exactly the main objective of our study: assessing measurement equivalence of PISA 2009 scales to see if there is in-equivalence in data. However, given the large number of countries and their cultural diversity, it is almost inevitable to find a certain level of in-equivalence in PISA scales. Hence, most likely the second research objective is of even greater importance because it examines the scope or degree of in-equivalence in data in light of its consequences with regard to country scores. As such, we go beyond mere warnings regarding cross-cultural comparisons of PISA test scores by indicating how much and in which way in-equivalence changes country scores.

Methodological Framework

Data

In this study, we used data from the newest wave of the PISA study—PISA 2009. The data set contains information about 475,460 15-year-old students coming from 64 countries or "economies." Student's performance in reading, mathematics, and science is assessed and estimated using the same test instruments translated into different languages. Original items were first formulated in English language and then translated into French. These two source questionnaires were then used for translation to all other languages.

PISA team divided the total set of questions into 13 linked testing booklets (PISA 2009—Assessment Framework). This means that each student was given a particular set of test items out of the complete item pool, which consequently led to a data set with large number of missing values. To avoid the issue of estimation of student's achievement levels on missing items and to

focus primarily on the issue of measurement equivalence, we decided to select only one booklet, that is, Booklet 8 for our analysis. Our choice of this particular booklet was made exclusively for practical reasons as Booklet 8 offered one of the shortest scales, thus providing an opportunity to conduct initial analyses with reduced computation time. However, since respondents were randomly assigned to one of the 13 booklets, results obtained in this booklet should be very similar to those in other booklets in the study and to the study as a whole.

There are 36,377 students from 64 countries/economies that were tested using Booklet 8. The three scales in Booklet 8 had 58 items in total, half of which was part of the "Reading" scale (29 items), whereas the other half was divided into "Mathematics" (12 items) and "Science" scales (17 items). Almost all items were scored as dichotomies, with categories "0 = no credit" for wrong or missing answers, and "1 = full credit" for correct answers. In addition, there are a few items that have an additional middle answer category ("partial credit") for partially correct answers, thus creating a 3-point ordinal item.[1]

Since nonanswers and multiple answers on the same question were scored as "0 = no credit," the remaining missing values represented test items that were not administered to the students and items that were deleted after assessment because of misprints or translation errors (OECD, 2009). There were relatively few of these missing values in the Mathematics and Science scales with percentage of missing values per item ranging from 0.3% to 6.5%. However, this percentage is larger in the Reading scale where it ranges from 0.2% to 21.5% with 12 items having more than 10% missing answers.[2] Because these values are missing by design, we assumed a random pattern of missingness (Little & Rubin, 1987) and applied a full information maximum likelihood procedure to use all available information from the data set.

Given that number of students per country was varying from 281 in Iceland to 2,935 in Mexico, we decided to weigh the cases to obtain a common group size of 600 per country/economy. In this way, the results would not be over- or underinfluenced by countries with higher or lower sample size, whereas the total sample size will remain approximately the same.

Method—Analysis of Differential Item Functioning

From a number of approaches developed for testing measurement equivalence of cross-cultural data, the most commonly used are multigroup confirmatory factor analysis (MCFA; Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000) and DIF analysis developed in the context of IRT (Raju, Laffitte, & Byrne, 2002). In addition, multigroup latent class analysis for models with categorical latent variables has been recently introduced (Kankaraš, Moors, & Vermunt, 2010; McCutcheon, 2002). The three approaches share a common core, that is, defining a measurement model by comparing the latent structure for several groups in a single model (Kankaraš, Vermunt, & Moors, 2011).

In this study, we decided to use IRT-based DIF approach to investigate the issue of measurement equivalence for a number of reasons. First, it seems that it fits the data best, given the fact that items are mostly dichotomies and that concepts measured, academic abilities in various domains are assumed to be normally distributed. Second, because IRT has been the dominant approach in the field of cognitive and educational testing, DIF analysis of measurement equivalence is the usual method of choice for these types of scales (Hambleton, 2001). Finally, PISA researchers themselves used IRT models (in particular "Rasch model") in their data analyses so that using the same modeling framework would add to the comparability of results.

At the heart of the IRT approach to measurement equivalence is the analysis of item bias, that is, DIF across groups (Raju et al., 2002; Thissen, Steinberg, & Wainer, 1988). DIF occurs when respondents from different countries with an equal position on the latent construct (e.g., equal knowledge or equal attitude) have different scores on the instrument items. This approach deals with nominal or ordinal response variables (both dichotomous and polytomous) while assuming continuous, normal distribution of latent constructs. It posits a nonlinear, logistic relationship between the latent construct and the observed score at the item level.

For dichotomous items, we used the 2-PL model (Birnbaum, 1968), whereas for ordinal items, we used the Generalized Partial Credit model for ordinal items (Muraki, 1992). These models differ from the Rasch model in that apart from the "difficulty" parameter, *b*, they also contain a "discrimination" parameter *a* for each item. The item discrimination (slope) parameter *a* is analogous to the factor loading and indicates how well items discriminate between subjects with different level of latent variable. For example, a math test item will have high discrimination if probability of answering it correctly steeply increases with increase of math skills and vice versa. The item difficulty (location) parameter *b* denotes the point on the latent trait at which the probability that a subject will endorse a given item in a given direction is 0.50. It denotes, for instance, how much mathematical knowledge a subject must possess to have a 50% probability of correctly answering a given item on a math test.

Multigroup formulations of these two models in which both "discrimination" and "difficulty" parameters are allowed to vary across groups are presented as follows:

$$\log\left[\frac{P(y_j=1\mid\Theta,g)}{P(y_j=0\mid\Theta,g)}\right]=a_j^g(\Theta-b_j^g), \tag{1}$$

$$\log\left[\frac{P(y_j=s\mid\Theta,g)}{P(y_j=s-1\mid\Theta,g)}\right]=a_j^g(\Theta-b_{js}^g), \tag{2}$$

for $2\le s\le S_j$, where s denotes one of the S_j categories of variable y_j, and Θ represents latent trait. Here, a_j^g is the discrimination parameter for group g and item j, b_j^g is the "difficulty" or location parameter for group g and item j, and b_{js}^g is the "difficulty" parameter for group g, item j, and category s. Hence, both the difficulty and discrimination parameters may vary across groups and cause in-equivalence or DIF. When DIF is present only in the location parameters b_j^g or b_{js}^g, it is called uniform DIF. Nonuniform DIF occurs when slope parameters a_j^g differ across groups (see Figure 1 for illustrations).

In this study, all analyses were carried out using Latent Gold 4.5 syntax module (Vermunt & Magidson, 2008). Latent Gold parameterizations of "difficulty" parameters differ from the models presented in Equations 1 and 2:

$$\log\left[\frac{P(y_{ij}=1\mid\Theta,g)}{P(y_{ij}=0\mid\Theta,g)}\right]=\tau_j^g-a_j^g\Theta_i, \tag{3}$$

$$\log\left[\frac{P(y_{ij}=s\mid\Theta,g)}{P(y_{ij}=s-1\mid\Theta,g)}\right]=\tau_{js}^g-a_{js}^g\Theta_i, \tag{4}$$

However, τ_j^g is just a rescaled version of b_j^g ($\tau_j^g=a_j^g b_j^g$) and as such has the same interpretation (Woods, Oltmanns, & Turkheimer, 2009).

Procedures in Analyzing Measurement Equivalence

In our analyses, we used a modified version of the likelihood ratio (LR) DIF test. LR DIF analysis involves a comparison of a number of IRT models varying in level of equivalence, that is, in constraints in variation of item parameter across groups. In particular, models that assume measurement equivalence and have both *a* and *b* parameters constrained to be equal across groups ("homogeneous" model) and models that assume both *a* and *b* parameters to vary freely ("heterogeneous" model) are compared in terms of their fit to the data with models that allow either difficulty or discrimination parameters to vary freely across groups (and thus, to be in-equivalent; Meade &

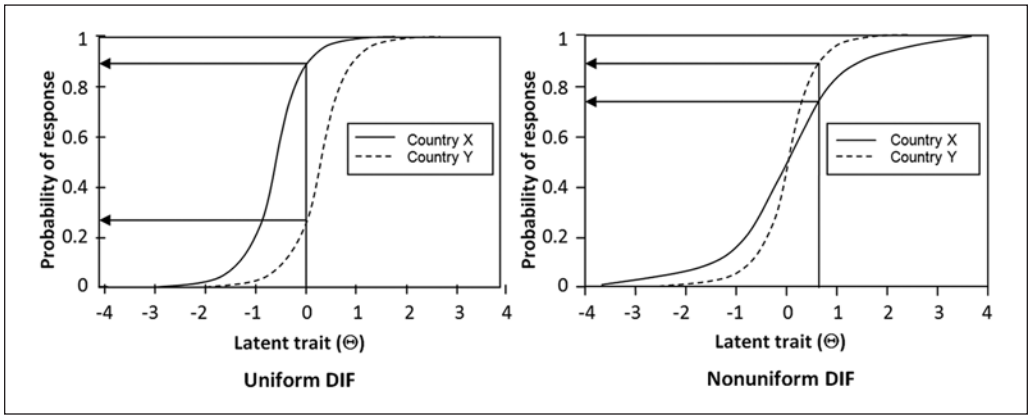

Figure 1. Item characteristic curves displaying uniform and nonuniform type of DIF.
Note. At equal levels of the underlying latent trait, respondents from the country X are estimated to have a higher probability of answering a question correctly than respondents from the country Y across all range of latent trait (Uniform DIF). At equal levels of the underlying latent trait, respondents from the countries X and Y are estimated to have different probabilities of answering a question correctly, with higher probabilities of respondents from country X when latent trait is smaller than 0 and vice versa (Nonuniform DIF). DIF = Differential Item Functioning.

Lautenschlager, 2004; Thissen et al., 1988). This is done separately for each item in a scale except for one reference or anchor item with the least amount of in-equivalence that is used to set a common scale for the latent variable and is for that reason constrained to be equal between groups.

We carried out two types of DIF tests per item. A first group of tests is analyzing nonuniform DIF in parameters and consists of comparison between two models (M_1 and M_2) that differ with regard to whether *a* parameters are permitted to vary across groups. In the heterogeneous model (M_1) they are free to vary, whereas in the second model (M_2) *a* parameters are set to be equal across groups. The *b* parameters are allowed to vary freely across groups in both models:

Nonuniform DIF test (DIF in *a* parameters): M_1: $a \neq a_g$ and $b \neq b_g$; versus M_2: $a = a_g$ and $b \neq b_g$.

The second group of tests examines uniform DIF by comparing the model with free *b* parameters and constrained *a* parameters (M_2) with the homogeneous model (M_3), whose *a* and *b* parameters are restricted to be equal across groups. Thus, the test of uniform DIF is a test of DIF in *b* parameters, conditional on the absence of DIF in *a* parameters:

Uniform DIF test (DIF in *b* parameters): M_2: $a = a_g$ and $b \neq b_g$; versus M_3: $a = a_g$ and $b = b_g$.

The analyses reported in this research include the comparison of 64 countries. Thus, model M_2 has 63 parameters more than homogeneous model (M_3)—one *b* parameter for a given item for each country, with one degree of freedom lost as they need to add up to their average (in case of a 3-point ordinal items M2 model has 126 extra parameters because there are two *b* parameters per item in each country). Similarly, the heterogeneous model (M_1) with both sets of parameters set free has 63 parameters more than model (M_2), in which a discrimination parameter of only one item at a time is constrained to be equal across all countries—one *a* parameter for a given item for each country.

Results

In the following sections, we first present outcomes from the analysis of uniform and nonuniform DIF as well as effects of item-level DIF on country scores for each scale separately. In the subsequent section, we show results in which we compare country scores from models that ignore DIF with those that account for it.

The LR chi-square (L^2) statistic is presented as a standard measure of model fit. However, when sample sizes are as large as it is the case in this study, LR tests tend to be too conservative, indicating misfit even for minimal differences between the two models. In addition, the LR statistic does not provide enough control for the number of parameters in a model that can sometimes be very large even for models of modest size (McCutcheon, 2002).

Due to these limitations of the LR statistic, we will also use Bayesian information criterion (BIC) as model fit criterion. BIC is based on LR but is designed to penalize models with larger numbers of parameters. Because more parameters in a model increase its likelihood, the information criteria reduce that likelihood by a certain amount that is a function of the increased number of estimated parameters. An additional benefit of using BIC compared with other information criteria such as AIC (Akaike information criterion) is that it also takes into account the sample size.[3] Thus, models with lower values of BIC indicate a better fit to the data, for a given number of parameters and given sample size.

We present changes in country scores to indicate the size of effects of the item-level DIF. In particular, we report how much country scores in particular scales change when comparing homogeneous models in which no DIF is allowed, with models in which DIF in only one item per time is allowed. These changes are presented in original scale units used in PISA studies (SD = 100) to make it easier to understand their practical relevance. Three measures are presented:

- Mean change in country scores across all 64 countries,
- Range of changes in country scores across countries, and
- Scaled change in country scores, which represent the product of first measure with the number of items in a scale.

The third measure is included because presented changes in country scores based on the item-level DIF depend on the number of items in a scale—the more items in a scale the smaller is the change in country scores, controlling for DIF size. Because three scales have different number of items per scale, this measure can be used to compare the amount of change in scales across scales. In addition, it can also be regarded as an indication of country scores changes on a scale level, assuming that all other items have the same DIF as the particular item in question.

Mathematics Scale

Analysis of measurement equivalence. Analyses of item DIF both for DIF in threshold parameters (uniform DIF–Table 1a) and in discrimination parameters (nonuniform DIF–Table 1b) are presented in Table 1. Results indicate that uniform DIF occurs in the Mathematics scale to a larger extent than nonuniform DIF. In particular, the change in LR chi-square value between compared models is much larger in case of uniform DIF relative to nonuniform DIF. When the LR chi-square change is calculated per added parameter this LR change/per parameter for uniform DIF is ranging from 6.0 until 35.8, with a mean value of 13.1, whereas the same change in case of nonuniform DIF varies only from as low as 2.2 until 8.6 with a mean value of 4.1. This means that most of the in-equivalence present in this scale is in the form of differences in thresholds rather than in differences in discrimination parameters. A similar conclusion can be reached from values of BIC statistic. In case of uniform DIF, there are eight DIF items (values written in bold), whereas none of the items has nonuniform DIF as measured by BIC. Presence of uniform DIF is especially large in the case of item "Third Side," and it is also larger in items "Diving-Q_1" and "Running Time-Q_1." In the remaining nine items, it does not vary much, with exception of the reference item that has substantially lower DIF.

Table 1. Model Fit Estimates and Items' DIF Sizes of the PISA "Mathematics" Scale.

Model	1a. DIF in threshold parameters (uniform DIF)			1b. DIF in discrimination parameters (nonuniform DIF)			1c. Size of uniform DIF in terms of changes in country scores per item		
	No of parameters	L^2	BIC (L^2)	No of parameters	L^2	BIC (L^2)	Mean change in country scores	Range of changes in country scores	Scaled changes in country scores
Homogeneous/heterogeneous[a]	90	116,180	−288,192	1,665	100,988	−286,759			
"View room" (33—Q_1)	153	115,593	**−288,115**[b]	1,602	101,132	−287,280	1.46	6.6	17.5
"Bricks" (34—Q_1)[c]	153	115,803	**−287,904**				1.15	4.2	13.8
"Pyramids" (155—Q_1)	153	115,478	**−288,229**	1,602	101,251	−287,161	2.09	6.5	25.0
"Pyramids" (155—Q_2)	153	115,349	**−288,358**	1,602	101,244	−287,168	2.71	10.5	32.5
"Pyramids" (155—Q_3)	153	115,567	**−288,140**	1,602	101,419	−286,992	1.50	5.4	17.9
"Pyramids" (155—Q_4)	153	115,587	**−288,120**	1,602	101,199	−287,213	1.33	5.1	15.9
"Diving" (411—Q_1)	153	115,197	**−288,510**	1,602	101,230	−287,182	2.24	7.1	26.8
"Diving" (411—Q_2)	153	115,409	**−288,298**	1,602	101,528	−286,884	1.16	5.8	13.9
"Braille" (442—Q_2)	153	115,383	**−288,325**	1,602	101,125	−287,287	1.81	5.0	21.7
"Third side" (462—Q_1)	153	112,117	**−291,590**	1,602	101,224	−287,187	2.62	21.9	31.4
"Running time"(474—Q_1)	153	115,271	**−288,436**	1,602	101,208	−287,204	1.65	6.3	19.7
"Labels" (803—Q_1)	153	115,503	**−288,205**	1,602	101,161	−287,251	1.59	5.8	19.1
Mean values							1.78	7.5	21.3

Note. DIF = Differential Item Functioning; PISA = Programme for International Student Assessment; BIC = Bayesian information criterion.

[a]In analysis of uniform DIF (Table 1a), the homogeneous model has 90 parameters, whereas in analysis of nonuniform DIF, the heterogeneous model has 1,665 parameters (Table 1b).

[b]BIC values that indicate item nonequivalence are presented in bold.

[c]This is the "anchor" item in this scale with fixed value of discrimination parameter ($a = 1$).

Analysis of uniform DIF effect sizes. In Table 1c, we present changes in country scores as function of uniform DIF in individual items, which offers insight into the practical effects of the uniform DIF that is found in the Mathematics scale. These changes are obtained by comparing country scores in the homogeneous model (no DIF) with country scores in models in which one item at a time is allowed to have uniform DIF.

First, in terms of mean changes in country scores, it seems that uniform DIF has a rather insubstantial effect, with an average mean change in country scores for all items and countries being 1.78. However, there is a substantial variation in this value both across items and across countries. The items "Pyramids Q_2" (2.71) and "Third Side Q_1" (2.62) have much larger effects than the reference item "Bricks Q1" (1.15). Variations are even larger across countries, where inspection of the range of changes in country scores reveals that Mathematics scores in some countries change for up to 4.2 points in reference item to as much as 21.9 points in the "Third Side Q_1" item.

Finally, when looking at the scaled changes in country scores, that is, when putting these item-level changed in country scores in the context of complete scales, we can see that the scaled change in country scores varies from 13.8 to 32.5, with an average for all items being 21.3 points. This measure is just a product of the first measure, mean change in country scores, with 12, which is the number of items in this scale. As such it can be interpreted as an expected country score change, assuming that all other items has same kind and size of DIF as the item in question. Because it is weighted with number of items in a scale, it can be used to compare changes in country scores per item across different scales.

"Science" Scale

Analysis of measurement equivalence. Similar to what was the case in the Mathematics scale, results in the Science scale indicate that the overwhelming part of the DIF in this scale is concentrated in threshold parameters in the form of uniform DIF (Table 2). In fact, this difference between uniform DIF and nonuniform DIF is even more pronounced in this scale, where the average amount of LR chi-square change per parameter is 16.5 in case of uniform DIF and only 4.1 for nonuniform DIF. Interestingly, however, the variation in uniform and nonuniform DIF across items is smaller in this scale, that is, DIF is more evenly spread across items than in the Mathematics scale. BIC statistics shows a similar pattern because 15 out of 17 items have uniform DIF as measured by BIC, whereas no nonuniform DIF was found in any item. In terms of individual items, the items "Green Parks Q_3" and "Development and Disaster Q_4" seem to be especially affected by uniform DIF, whereas anchor item "Bacteria in Milk Q_3" and "Different Climates Q_2" are least affected.

Analysis of uniform DIF effect sizes. The mean changes in country scores per individual item are slightly smaller than in the Mathematics scale, with an average of 1.4 points, the minimum being 0.8 for the item "Different Climates Q_4" and the maximum 2.1 for the item "Different Climates Q_1." Variation in score changes across countries is also smaller than in case of the Mathematics scale, ranging from maximum of 2.6 ("Different Climates Q2") until 15.1 ("Development and Disaster Q4").

However, when looking at scaled changes in country scores, we can see that DIF effects are actually more substantial in this scale than in the Mathematics scale, with average scaled changes in country scores for all items being 23.9. This indicates that although country scores between item-level DIF and homogeneous models in average change less in Science scale, they are actually more substantial taking into account larger number of items in this scale (17 items) compared with the Mathematics scale.

Table 2. Model Fit Estimates for Various Multigroup Models of the PISA "Science" Scale.

Model	2a. DIF in threshold parameters (uniform DIF)			2b. DIF in discrimination parameters (non-uniform DIF)			2c. Size of uniform DIF in terms of changes in country scores per item		
	No. of parameters	L^2	BIC (L^2)	No. of parameters	L^2	BIC (L^2)	Mean change in country scores	Range of changes in country scores	Scaled changes in country scores
Homogeneous/heterogeneous[a]	98	288,342	−115,945	2,177	266,904	−115,438			
"Good vibrations" (131—Q_2)	161	287,255	**−116,368**[b]	2,114	267,120	−115,888	1.7	5.8	28.3
"Good vibrations" (131—Q_4)	161	287,334	**−116,289**	2,114	267,152	−115,855	1.3	6.6	21.3
"Solar power" (415—Q_2)	161	287,502	**−116,121**	2,114	267,212	−115,796	1.1	7.0	18.3
"Solar power" (415—Q_7)	161	287,148	**−116,475**	2,114	267,199	−115,808	1.2	3.4	19.7
"Solar power" (415—Q_8)	161	287,537	**−116,086**	2,114	267,115	−115,892	1.1	3.8	18.0
"Bacteria in milk" (428—Q_1)	161	287,255	**−116,367**	2,114	267,043	−115,964	1.5	5.9	24.9
"Bacteria in milk"(428—Q_3)[c]	161	287,816	−115,806				1.2	4.3	20.2
"Bacteria in milk" (428—Q_5)	161	287,372	**−116,251**	2,114	267,109	−115,898	1.6	6.3	28.0
"Green parks" (438—Q_1)	161	287,101	**−116,522**	2,114	267,111	−115,896	1.5	6.5	25.1
"Green parks" (438—Q_2)	161	287,669	**−115,954**	2,114	267,221	−115,787	1.1	3.5	18.0
"Green parks" (438—Q_3)	161	286,639	**−116,984**	2,114	267,122	−115,885	1.7	6.8	29.8
"Different climates" (465—q_1)	161	287,111	**−116,511**	2,114	267,136	−115,871	2.1	5.3	35.7
"Different climates" (465—Q_2)	161	287,728	−115,895	2,114	267,303	−115,705	0.9	2.6	16.0
"Different climates" (465—Q_4)	161	287,353	**−116,270**	2,114	267,331	−115,676	0.8	2.7	13.0
"Dev. and disaster"(514—Q_2)	161	287,027	**−116,596**	2,114	267,166	−115,842	2.0	5.3	34.0
"Dev. and disaster" (514—Q_3)	161	287,319	**−116,304**	2,114	267,159	−115,848	1.3	5.2	22.5
"Dev. and disaster'"(514—Q_4)	161	286,637	**−116,986**	2,114	267,104	−115,904	2.0	15.1	33.6
Mean values							1.4	5.7	23.9

Note. PISA = Programme for International Student Assessment; DIF = Differential Item Functioning; BIC = Bayesian information criterion.

[a]In analysis of uniform DIF (Table 2a), homogeneous model has 98 parameters, whereas in analysis of nonuniform DIF, heterogeneous model has 2,177 parameters (Table 2b).

[b]BIC values that indicate item nonequivalence are presented in bold.

[c]This is the "anchor" item in this scale with fixed value of discrimination parameter ($a = 1$).

Reading Scale

Analysis of measurement equivalence. In the case of the Reading scale LR chi-square differences between models very similar to those in the Science scale again indicating that the vast majority of DIF present in the scale is in the form of uniform DIF (Table 3). This can be best seen in terms of LR chi-square change per parameter—16.4 compared with 4.3 on average across items, which is very similar to the values obtained in the Science scale. However, these values vary more than in the previous scale and are in that regard more similar to the Mathematics scale.

BIC statistics also point out that most of the items have significant uniform DIF (24 of 29), whereas none of them exhibits nonuniform DIF, which is in line with results from the other two scales. "Optician Q_1" item has the biggest DIF in terms of LR chi-square statistic (L^2/par of 36.2), whereas the smallest DIF occurs with anchor item "Childs Futures Q_{10}" (L^2/par of 7.0) and "Languages Q_5" (L^2/par of 8.5).

Analysis of uniform DIF effect sizes. Compared with the two previous scales the mean changes in country scores per item in Reading scales are smaller both in terms of its absolute values as well as in terms of their range in variation across items and countries. Across items, they vary from as little as 0.1 ("Languages Q6") to 1.6 ("Optician Q2"), with an average of 0.9. Maximum shifts in country scores per item range from 0.4 ("Languages Q6") to 7.9 ("Telephone Q1").

Scaled changes in country scores, however, are of much more substantial value, because they are weighted with 29, the number of items in this scale. Thus, they range from 2.5 ("Languages Q6") to 46 points ("Optician Q2"), with an average of 25.0 representing the highest effect of DIF on country scores among the three scales.

Consequences of Measurement In-Equivalence at Scale Level

Results presented so far clearly indicate existence of DIF in PISA 2009 data set. As much as these results may be important on they own, when working with empirical datasets it is worthwhile to transmit the results of ME analysis into the concrete research situation and delineate their implications and effects. The question is then what are the consequences of these findings from a practical point of view. In other words, it is important to see what these results imply in terms of main products of PISA data sets—country scores and consecutive country comparisons.

In following paragraphs, we will compare country scores between two models. The first one is the homogeneous model that assumes that there is no in-equivalence in the data sets, that is, all parameters are the same across countries. This homogeneous model is compared with the model which assumes presence of uniform DIF in all but anchor items, that is, it allows threshold parameters to vary freely across countries. Note that discrimination parameters of the second model are set to be equal across countries. Thus, by comparing the country scores between these two models, we are able to see practical consequences of uniform DIF expressed in original PISA scale units (with $SD = 100$).

In Figures 2, 3, and 4, we present these differences in country scores between the homogeneous (equivalent) and uniform DIF (in-equivalent) model. Values shown represent how many points score of a given country decreased (negative values), remain the same (values around 0), or increased (positive values) from the homogeneous model to the in-equivalent model. In other words, presented values can be seen as indication of the amount of bias in those country scores that do not account for measurement equivalence.

As far as the Mathematics scale is concerned (Figure 2), differences in country scores ranging from 0 (Spain) to 43 (Peru), with average difference of 13.8 points. Around half of the countries (33) have relatively unsubstantial difference of less than 10 points between two models. However, differences in the other half of the countries are more pronounced and in some cases rather substantial. For example, after accounting for DIF, country scores in Shanghai decrease for 40 and

Table 3. Model Fit Estimates for Various Multigroup Models of the PISA "Reading" Scale.

	3a. DIF in threshold parameters (uniform DIF)			3b. DIF in discrimination parameters (nonuniform DIF)			3c. Size of uniform DIF in terms of changes in country scores per item		
Model	No. of parameters	L^2	BIC (L^2)	No. of parameters	L^2	BIC (L^2)	Mean change in country scores	Range of changes in country scores	Scaled changes in country scores
Homogeneous/heterogeneous[a]	126	804,737	400,745	3,969	762,497	399,072			
"Drugged spiders" (55—Q_1)	189	803,916	**400,589**[b]	3,906	762,833	398,743	0.7	2.5	18.8
"Drugged spiders" (55—Q_2)	189	803,558	**400,230**	3,906	762,788	398,697	0.7	4.9	21.0
"Drugged spiders" (55—Q_3)	189	803,122	**399,794**	3,906	762,731	398,641	1.3	3.8	36.3
"Drugged spiders" (55—Q_5)	189	803,900	**400,573**	3,906	762,705	398,615	0.8	4.1	24.0
"Telephone" (104—Q_1)	189	803,868	**400,541**	3,906	762,704	398,613	0.8	7.9	23.9
"Telephone" (104—Q_2)	189	803,401	**400,074**	3,906	762,786	398,696	0.5	2.3	15.1
"Telephone" (104—Q_5)	189	802,893	**399,566**	3,906	762,760	398,669	1.1	3.9	31.6
"Exchange" (111—Q_1)	189	803,584	**400,256**	3,906	762,818	398,727	1.0	2.8	29.1
"Exchange" (111—Q_2)	189	802,820	**399,493**	3,906	762,728	398,638	1.4	5.7	40.5
"Exchange" (111—Q_6)	189	803,141	**399,814**	3,906	762,783	398,693	1.3	5.0	37.3
"Optician" (227—Q_1)	189	802,456	**399,129**	3,906	762,706	398,616	1.1	3.5	31.0
"Optician" (227—Q_2)	189	802,325	**398,998**	3,906	762,832	398,741	1.6	6.1	46.0
"Optician" (227—Q_3)	189	803,478	**400,151**	3,906	762,742	398,651	1.0	5.6	29.8
"Optician" (227—Q_6)	189	803,595	**400,267**	3,906	762,803	398,713	1.2	4.0	34.8
"Languages" (412—Q_1)	189	803,944	**400,617**	3,906	762,830	398,740	0.6	2.2	16.6
"Languages" (412—Q_5)	189	804,204	400,877	3,906	762,952	398,862	0.5	1.5	14.2
"Languages" (412—Q_6)	189	804,048	**400,721**	3,906	762,651	398,561	0.1	0.4	2.5
"Languages" (412—Q_8)	189	803,990	**400,663**	3,906	762,639	398,549	0.9	4.3	25.2
"Childs' futures" (420—Q_2)	189	803,921	**400,594**	3,906	762,739	398,648	0.9	3.2	26.9
"Childs' futures" (420—Q_6)	189	804,118	400,790	3,906	762,767	398,677	0.6	2.5	16.8
"Childs' futures" (420—Q_9)[c]	189	804,296	400,969				0.5	1.7	13.6
"Childs' futures" (420—Q_{10})[a]	189	803,884	**400,558**	3,906	762,754	398,663	1.2	6.5	34.3
"Narcissus" (437—Q_1)	189	803,939	**400,612**	3,906	762,738	398,648	0.4	1.6	12.5
"Narcissus" (437—Q_6)	189	803,594	**400,267**	3,906	762,792	398,701	0.6	2.1	17.1
"Narcissus" (437—Q_7)	189	804,090	400,763	3,906	762,708	398,618	0.5	2.1	14.4
"Summer job" (453—Q_1)	189	804,132	400,805	3,906	762,818	398,728	0.7	4.0	20.2
"Summer job" (453—Q_4)	189	803,863	**400,536**	3,906	762,864	398,774	0.8	2.8	23.6
"Summer job" (453—Q_5)	189	803,855	**400,527**	3,906	762,743	398,653	0.9	5.5	26.2
"Summer job" (453—Q_6)	189	803,625	**400,298**	3,906	762,773	398,682	1.1	3.8	31.4
Mean values							0.9	3.7	25.0

Note. PISA = Programme for International Student Assessment; DIF = Differential Item Functioning; BIC = Bayesian information criterion.

[a]In analysis of uniform DIF (Table 3a), homogeneous model has 126 parameters, whereas in analysis of nonuniform DIF, heterogeneous model has 3,969 parameters (Table 3b).

[b]BIC values that indicate item nonequivalence are presented in bold.

[c]This is the "anchor" item in this scale with fixed value of discrimination parameter ($a = 1$).

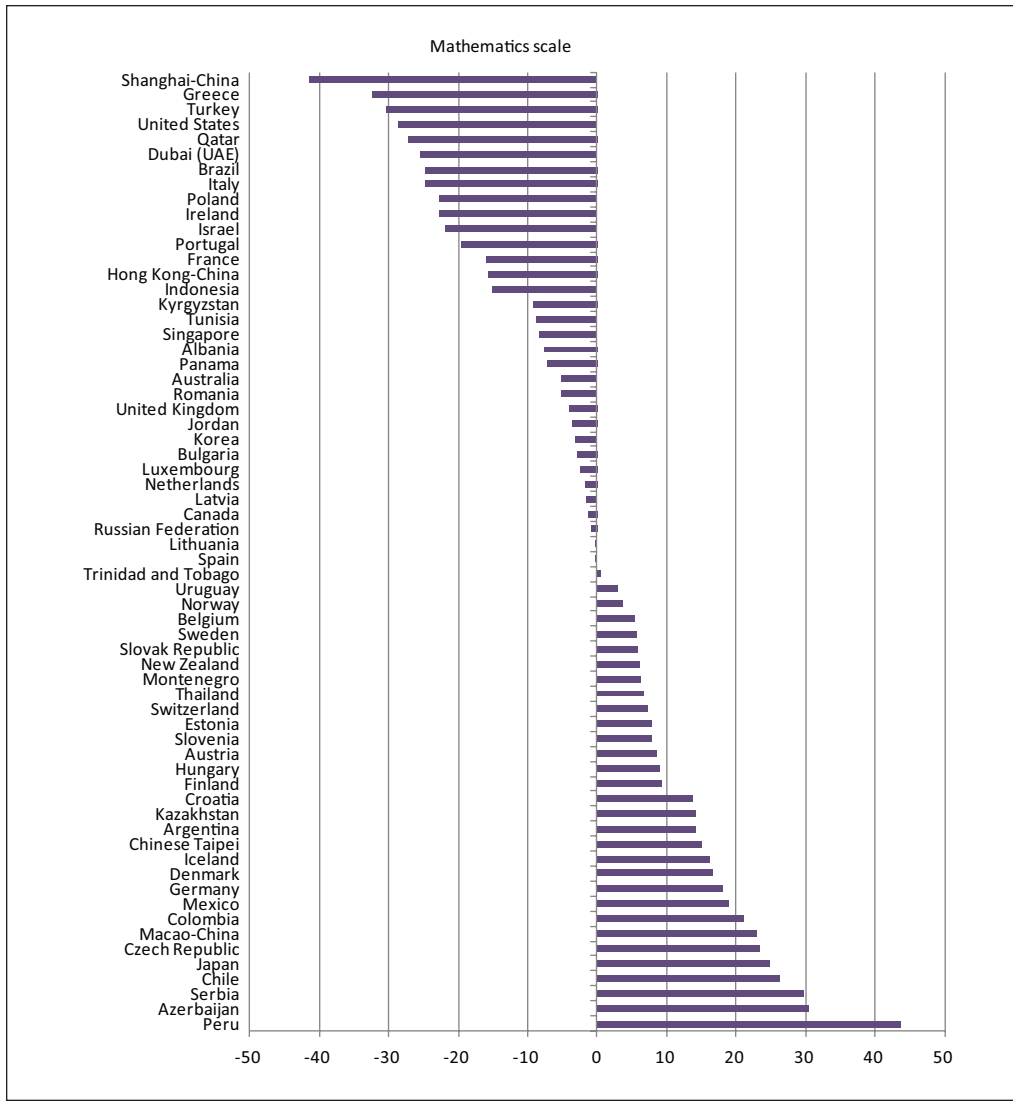

Figure 2. Differences between country scores in PISA "Mathematics" scale when comparing homogeneous model (M_1: *same difficulty and discrimination parameters across countries*) with uniform DIF model (M_2: *different difficulty and same discrimination parameters across countries*): difference = $M_1 - M_2$.
Note. standard deviation of both scales is 100. PISA = Programme for International Student Assessment; DIF = Differential Item Functioning.

in Greece, Turkey, the United States, and so on for about 30 points from the homogeneous model. At the same time, country scores increase for more than 40 points in Peru and about 30 points in Azerbaijan, Serbia, Chile, and so on. However, it should be noted that these changes are not drastic, and that most countries have similar rankings in the new models as well.

Country scores in the Science scale differ slightly more across the two models than what was the case in the Mathematics scale (Figure 3). Here, allowing for uniform DIF within the inequivalent model brings upon average changes in country scores of 14.8 points, ranging from 0 in France to 41 points in Singapore. There are 29 countries with differences of less than 10 points, and 18 countries with differences larger than 20 points. Change in scores is particularly large in

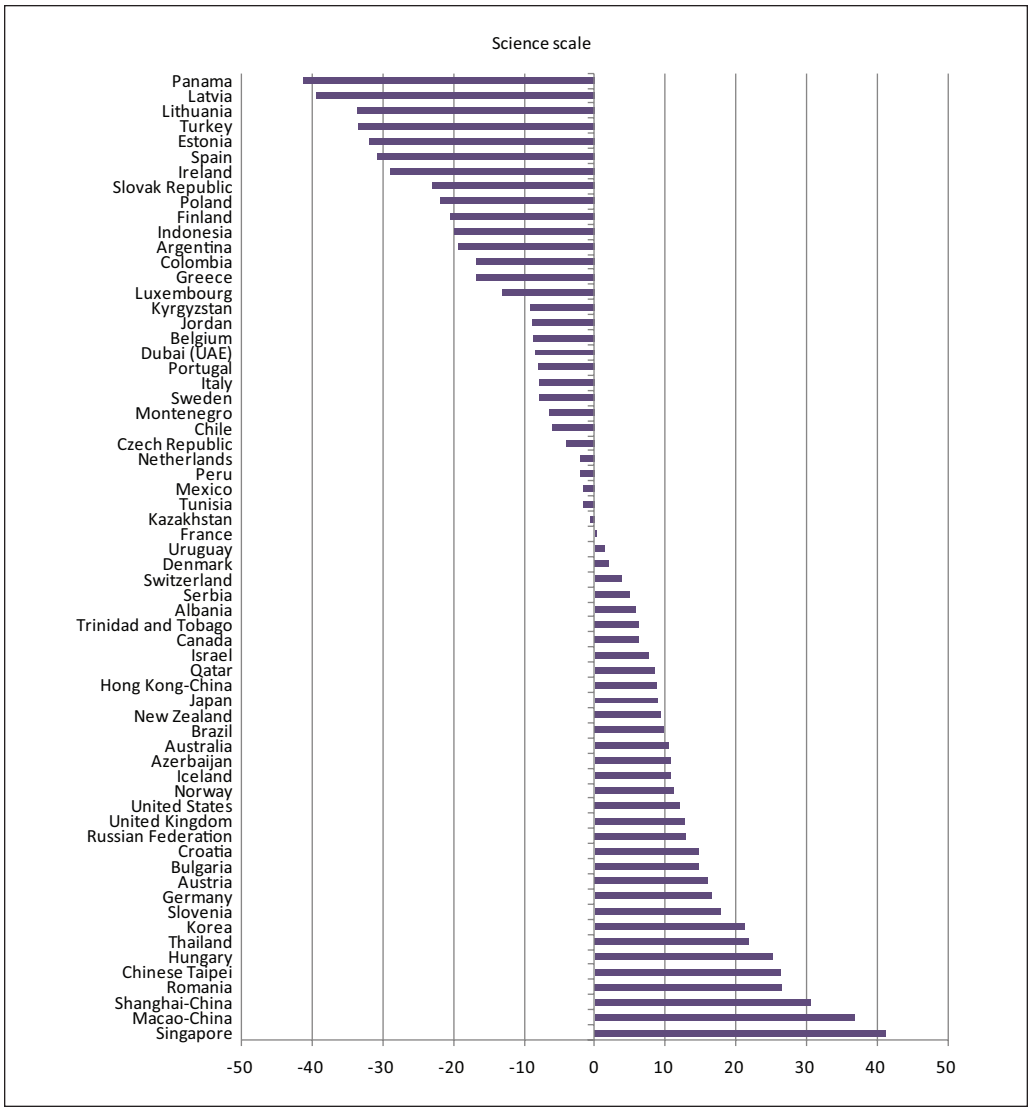

Figure 3. Differences between country scores in PISA "Science" scale when comparing homogeneous model (M_1: *same difficulty and discrimination parameters across countries*) with uniform DIF model (M_2: *different difficulty and same discrimination parameters across countries*): difference = $M_1 - M_2$.
Note. standard deviation of both scales is 100. PISA = Programme for International Student Assessment; DIF = Differential Item Functioning.

case of Panama, Latvia, Lithuania, Turkey, Estonia, and Spain whose scores decrease for more than 30 points. In contrast, scores for Singapore, Macao, and Shanghai increase for more than 30, and for Romania, Chinese Taipei, Hungary, Thailand, and Korea for more than 20 points.

Finally, differences in country scores on the Reading scale are the most pronounced among all three scales (Figure 4). Differences vary from 0 points in Brazil to 66 points in Macao-China, with average being almost 20 points (19.1). Only 19 countries have differences of less than 10 points, with 22 of them differing between models for more than 20 points. In particular, country scores substantially decreased in Norway (57 points), Denmark (46), Germany (39), Israel (34), and Poland (33). However, they strongly increased in Macao (66), Turkey (52), Serbia (48), Montenegro (46), Chinese Taipei (43), Russia (34), Belgium (31), and Shanghai-China (31).

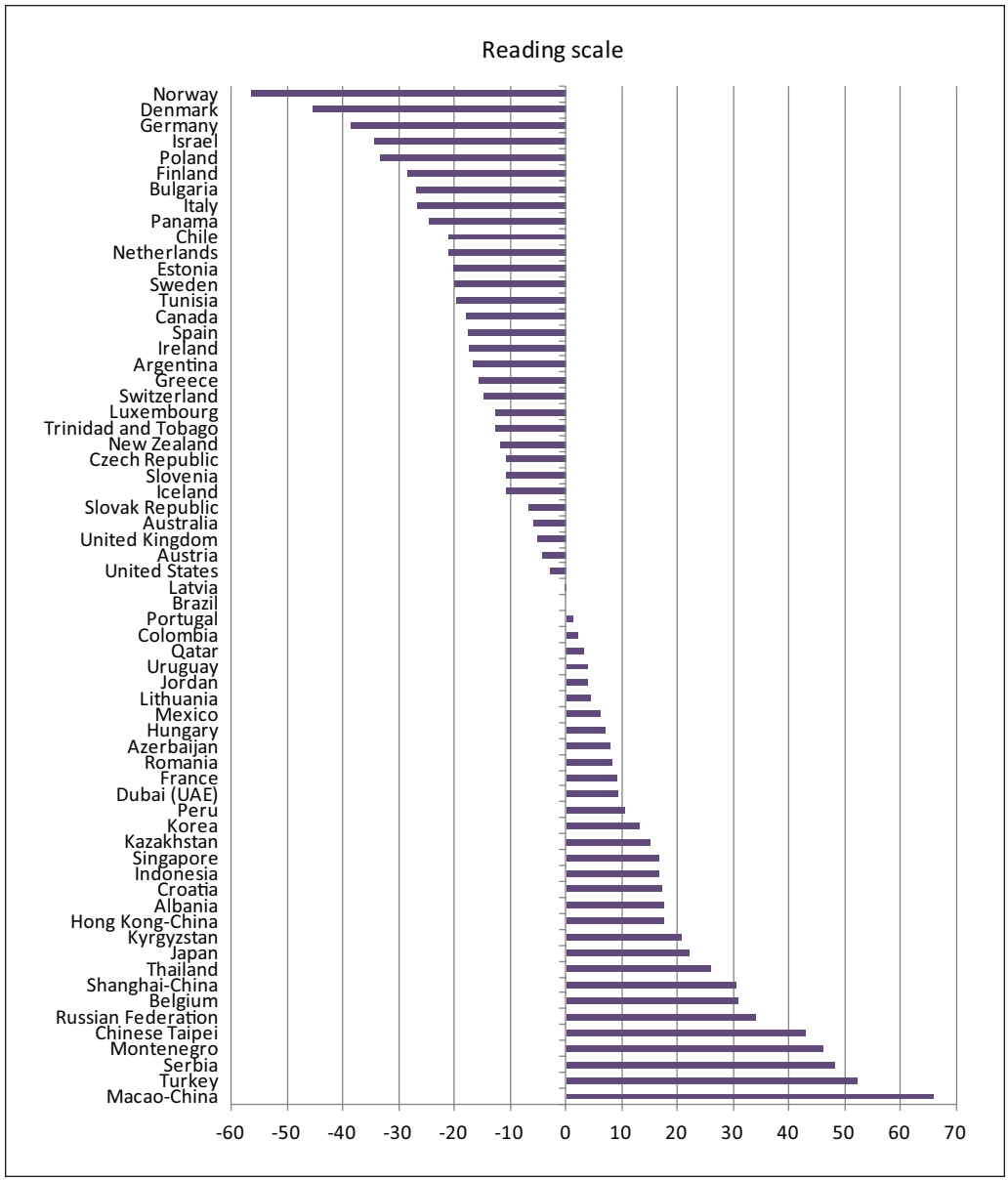

Figure 4. Differences between country scores in PISA "Reading" scale when comparing homogeneous model (M_1: *same difficulty and discrimination parameters across countries*) with uniform DIF model (M_2: *different difficulty and same discrimination parameters across countries*): difference = $M_1 - M_2$.
Note. standard deviation of both scales is 100. PISA = Programme for International Student Assessment; DIF = Differential Item Functioning.

Discussion

Results presented in the previous sections offer a number of noteworthy observations. The most important information that can be deducted from them is the fact that the PISA 2009 scales are comprised of items most of which are in-equivalent. This may come as a surprise given the considerable effort and numerous methodological safeguards that members of PISA teams have used in the process of construction and implementation of PISA tests across countries. More importantly, it

looks especially at odds with team's emphasis on comparability of country scores and validity of the prime product of PISA studies—its "league tables." However, these results are less surprising if one takes into account that these scales are translated and administered in the large number of countries that are coming from a very different cultural and linguistic backgrounds. Furthermore, obtained results are in accordance with research on equivalence of previous versions of PISA, which also indicated substantial presence of in-equivalence (Grisay & Monseur, 2007; Hopmann & Brinek, 2007).

Another important observation is that the major part of in-equivalence in the PISA 2009 data set is located in difficulty parameters (uniform DIF), rather than in the discrimination parameters (nonuniform DIF). This means that those factors that influenced results and caused DIF are not associated with measured academic competencies. This result is expected because equivalence in difficulty parameters implies absence of any other factor with different effects on results between countries, and as such represent stricter and harder to obtain forms of equivalence.

Although each item was affected by DIF to a certain degree, there was a substantial amount of variation in changes of country score as a consequence of the item-level DIF. Having access to the exact wording of the items could shed a light on what might be the cause of these differences but this is unfortunately not possible because PISA questionnaires are not available for research purposes. As expected, changes in country scores are the most substantial in the Mathematics scale and the least substantial in the Reading scale because they depend on the total number of items in a scale—the more items in a scale the smaller is the effect of DIF in each of the individual items on scale score. Although these changes at first may seem marginal, they still make an important influence on the final scores, due to the fact that most of the items in the three scales are in-equivalent and that, as a consequence, DIF accumulates at the scale level. However, what is noticeable on the scale level is that there is greater in-equivalence in the "Science" and especially the Reading scale, compared with the Mathematics scale. This is indicated by much smaller proportion of DIF items judged by BIC statistic and smaller values of scaled changes in country scores, as well as in comparison of country scores between the homogeneous and the in-equivalent models. In these comparisons it became evident that uniform DIF in the Mathematics scale seems to have less influence on country scores than what is the case in the other two scales. Whereas mathematics, but also science, has a somewhat "universal" aspect of language (sharing formulas and symbols), the fact that the "reading test" shows the highest difficulty in establishing measurement equivalence indicates that language specific characteristics are difficult to translate in an universal test.

When inspecting changes in country scores across the homogeneous and the in-equivalent model, a peculiar pattern can be noticed. In particular, it is indicative to see that countries with similar linguistic and cultural background tend to have same type of shift in a given scale. For example, after accounting for in-equivalence by allowing *b* parameters of DIF items to vary freely, country scores in Science scale increased in almost all of Southeast Asian countries/territories—Singapore, Macao-China, Shanghai-China, Chinese Taipei, and Korea. At the same time, scores decrease in all three Baltic countries—Lithuania, Latvia, and Estonia. Similar patterns occur in the case of Reading scale; country scores again increase in Southeast Asian countries—Macao-China, Shanghai-China, Chinese-Taipei—as well as in the two culturally and linguistically close countries, Serbia and Montenegro. However, scores decrease in the group of the Scandinavian countries, namely, Norway, Denmark, Finland, and Sweden. This pattern indicates that factors that cause in-equivalence—whether it is a language, cultural background, educational system, or something else—work on a regional level, rather than (or as well as) on the level of individual countries. In other words, it seems that the in-equivalence often occurs in the form of differences in regional characteristics that affect student's answers and not only in the form of particular country-specific factors that are independent from given geo-political and cultural context of a particular country. This also further implies that countries within the same regions are more likely to have equivalent results when compared with each other than when

compared with countries that are not part of that particular regional group. This is in line with the principle of cultural proximity that states that cultural groups that are closer to each other, in the cultural, historical, and linguistic sense, are also more likely to have equivalent, that is, comparable results (Hui & Triandis, 1985; Kankaraš & Moors, 2012).

Conclusion

Whenever a cross-cultural or cross-national comparison is intended, it is necessary to first check if or to what degree results are comparable. But this requirement is even more important in the case of PISA scales because PISA is not simply a testing regime. PISA is constructed and motivated under a clear and specific policy framework to "provide high quality data to support policy formation and review" (OECD, 2003, p. 3). Indeed, PISA data are used to justify change or provide support for existing policy direction in both the domestic and the European contexts in the way that few other international studies in education ever did (Grek, 2009; Hopmann & Brinek, 2007; Mortimore, 2009; Pongratz, 2006). Thus, consequences of invalidity of its data can be very serious and wide-ranging.

This study reveals the presence of in-equivalence in PISA 2009 data set in the form of uniform DIF. In-equivalence occurred in a majority of test questions in all three scales researched and is, on average, of moderate size. It varies considerably both across items and across countries. When this uniform DIF is accounted for in the in-equivalent model, the resulting country scores change considerably in the cases of the "Mathematics," "Science," and especially, Reading scale. These changes tend to occur simultaneously and in the same direction in groups of regional countries. The most affected seems to be Southeast Asian countries/territories whose scores, although among the highest in the initial, homogeneous model, additionally increase when accounting for in-equivalence in the scales.

Presented results clearly illustrate the importance of the analysis of measurement equivalence in PISA 2009 data set, indicating size and scope of in-equivalence in the data set. But most importantly, they indicate that presence of in-equivalence impairs the validity of country comparisons and reduces their accuracy. It is therefore necessary to take this uncertainty into account when interpreting PISA "league tables." Instead of assigning fundamental importance to all the minute differences in the original country scores, a researcher should make sure that in-equivalence is accounted for in measurement models, and explicitly incorporated in the interpretation of the country scores and their comparisons.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. There was also an additional answer category "not reached" for all consecutive missing values clustered at the end of test session, except for the first value of the missing series, which is coded as item level nonresponse. However, since there are very few answers in this category (less than 0.5% in most items), to simplify the analyses, we decided to merge these answers with category "0" indicating nonresponse.
2. In the "Mathematics" scale, one item was not administered at all in three countries, whereas in the "Reading" scale, seven countries had 1, 2, and in one case, 4 not-administered items.
3. BIC = L^2 − df*[ln(N)]; AIC = L^2 − 2df.

References

Allerup, P. (2007). Identification of group differences using PISA scales—Considering effects of inhomogeneous items. In S. Hopmann & G. Brinek (Eds.), *PISA according to PISA* (pp. 175-201). Wien, Austria: Lit-Verlag, University of Vienna.

Bechger, T. M. (2006, June). *On comparative validity*. Paper presented at the ICT conference in Brussels, Belgium.

Birnbaum, A. (1968). Some latent trait models. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-479). Reading, MA: Addison-Wesley.

Bonnet, G. (2002). Reflections in a critical eye: On the pitfalls of international assessment. *Assessment in Education*, *9*, 387-400.

Fertig, M. (2004). *What can we learn from international student performance studies? Some methodological remarks*, RWI: Discussion Paper No. 23, Rheinisch-Westfälisches Institut für Wirtschaftsforschung.

Goldstein, H. (2008). Comment peut-on utiliser les etudes comparatives internationale pour doter les politiques educatives d'informations fiables? *Revue Francaise de Pedagogie*, *164*, 69-76.

Grek, S. (2009). Governing by numbers: The PISA "effect" in Europe. *Journal of Educational Policy*, *24*, 23-27.

Grisay, A., & Monseur, C. (2007). Measuring the equivalence of item difficulty in the various versions of an international test. *Studies in Educational Evaluation*, *33*, 69-86.

Hambleton, R. K. (2001). The next generation of the ITC test translation and adaptation guidelines. *European Journal of Psychological Assessment*, *17*, 164-172.

Hambleton, R. K., & Kanjee, A. (1993, April). *Enhancing the validity of cross-cultural studies: Improvements in instrument translation methods*. Paper presented at the meetings of AERA and NCME national studies, Atlanta, GA.

Hambleton, R. K., Merenda, P., & Spielberger, C. (2005). *Adapting educational and psychological tests for cross cultural assessment*. Mahwah, NJ: Lawrence Erlbaum.

Harkness, J. A., Pennell, B. E., & Schoua-Glusberg, A. (2004). Survey questionnaire translation and assessment. In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, & E. Singer (Eds.), *Methods for testing and evaluating survey questionnaires* (pp. 435-473). Hoboken, NJ: John Wiley & Sons.

Hopmann, S. T., & Brinek, G. (2007). PISA according to PISA—Does PISA keep what it promises? In T. Hopmann, G. Brinek, & M. Retzl (Eds.), *PISA zufolge PISA–PISA according to PISA* (pp. 9-19). Wien, Austria: Lit-Verlag.

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, *18*, 117-144.

Hui, C. H., & Triandis, H. C. (1985). Measurement in cross-cultural psychology: A review and comparison of strategies. *Journal of Cross-cultural Psychology*, *16*, 131-152.

Hui, C. H., & Triandis, H. C. (1989). Effects of culture and response format on extreme response style. *Journal of Cross-Cultural Psychology*, *20*, 296-309.

Kankaraš, M., & Moors, G. (2010). Researching measurement equivalence in cross-cultural studies. *Psihologija*, *43*, 121-136.

Kankaraš, M., & Moors, G. (2012). Cross-national and cross-ethnic differences in political and leisure attitudes: A case of Luxemburg. *Cross-Cultural Research*, *46*, 224-254.

Kankaraš, M., Moors, G., & Vermunt, J. K. (2010). Testing for measurement invariance with latent class analysis. In E. Davidov, P. Schmidt, & J. Billiet (Eds.), *Cross-cultural analysis: Methods and applications* (pp. 359-384). New York, NY: Routledge.

Kankaraš, M., Vermunt, J. K., & Moors, G. (2011). Measurement equivalence of ordinal items: A comparison of factor analytic, item response theory, and latent class approaches. *Sociological Methods & Research*, *40*, 279-310.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York, NY: John Wiley & Sons.

McCutcheon, A. (2002). Basic concepts and procedures in singe-and multiple-group latent class analysis. In J. Hagenaars & A. McCutcheon (Eds.), *Applied latent class analysis* (pp. 56-88). Cambridge, UK: Cambridge University Press.

Meade, A. W., & Lautenschlager, G. J. (2004). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods*, *7*, 361-388.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research*, *13*, 127-143.

Mortimore, P. (2009). *Alternative models for analysing and representing countries' performance in PISA*. Education International, Brussels (http://download.ei-ie.org/Docs/WebDepot/Alternative%20Models%20in%20PISA.pdf) .

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159-177.

OECD. (1999). *Measuring student knowledge and skills: A new framework for assessment*. Paris, France: OECD Publishing.

OECD. (2001). *Knowledge and skills for life: First results from programme for international student assessment*. Paris, France: OECD Publishing.

OECD. (2003). *PISA 2003 technical report*. Paris, France: OECD Publishing.

OECD. (2009). *PISA data analysis manual: SPSS and SAS* (2nd ed.). Paris, France: OECD Publishing.

OECD. (2010). *PISA 2009 results*. Paris, France: OECD Publishing.

Pongratz, L. (2006). Voluntary self-control: Education reform as a governmental strategy. *Education Philosophy and Theory*, *38*, 471-482.

Poortinga, Y. H. (1989). Equivalence of cross-cultural data: An overview of basic issues. *International Journal of Psychology*, *24*, 737-756.

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology*, *87*, 517-529.

Sjøberg, S. (2007). PISA and "real life challenges": Mission impossible? In S. T. Hopmann, G. Brinek, & M. Retzl (Eds.), *PISA zufolge PISA–PISA according to PISA* (pp. 203-224). Vienna, Austria: Lit-Verlag.

Steenkamp, J. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, *25*, 78-90.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 147-169). Hillsdale, NJ: Lawrence Erlbaum.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *2*, 4-69.

Van de Vijver, F. (1998). Towards a theory of bias and equivalence. *Zuma Nachrichten: Cross-Cultural Survey Equivalence*, *3*, 41-65.

Van de Vijver, F., & Poortinga, Y. H. (1991). Testing across cultures. In R. K. Hambleton & J. Zaal (Eds.), *New developments in testing: Theory and applications* (pp. 277-308). Dordrecht, The Netherlands: Kluwer.

Vermunt, J. K., & Magidson, J. (2008). *LG-syntax user's guide: Manual for latent GOLD 4.5 syntax module*. Belmont, MA: Statistical Innovations.

Woods, C. M., Oltmanns, T. F., & Turkheimer, E. (2009). Illustration of MIMIC-model DIF testing with the schedule for nonadaptive and adaptive personality. *Journal of Psychopathology and Behavioral Assessment*, *31*, 320-330.

Yildirim, H. H. (2006). *The differential item functioning (DIF) analysis of mathematics items in the international assessment programs* (Unpublished doctoral thesis). Middle East Technical University, Ankara, Turkey.